J Comput Aided Mol Des (2009) 23:355–374
DOI 10.1007/s10822-009-9266-3

Effects of protein conformation in docking: improved pose prediction through protein pocket adaptation

Ajay N. Jain

Received: 23 January 2009 / Accepted: 14 March 2009 / Published online: 2 April 2009

Abstract Computational methods for docking ligands have been shown to be remarkably dependent on precise protein conformation, where acceptable results in pose prediction have been generally possible only in the artificial case of re-docking a ligand into a protein binding site whose conformation was determined in the presence of the same ligand (the "cognate" docking problem). In such cases, on well curated protein/ligand complexes, accurate dockings can be returned as top-scoring over 75% of the time using tools such as Surflex-Dock. A critical application of docking in modeling for lead optimization requires accurate pose prediction for novel ligands, ranging from simple synthetic analogs to very different molecular scaffolds. Typical results for widely used programs in the "cross-docking case" (making use of a single fixed protein conformation) have rates closer to 20% success. By making use of protein conformations from multiple complexes, Surflex-Dock yields an average success rate of 61% across eight pharmaceutically relevant targets. Following docking, protein pocket adaptation and rescoring identifies single pose *families* that are correct an average of 67% of the time. Consideration of the *best of two* pose families (from alternate scoring regimes) yields a 75% mean success rate.

Keywords Docking · Cross-docking · Protein flexibility · Pose prediction · Surflex · Surflex-Dock

A. N. Jain (✉)
Department of Bioengineering and Therapeutic Sciences,
University of California, San Francisco, CA 94158-9001, USA
e-mail: ajain@jainlab.org

Introduction

The field of molecular docking for the purpose of small molecule drug design is relatively mature. The 1980s saw the establishment of the field with the pioneering work of Blaney and Kuntz on rigid docking of small molecules to protein structures [1]. The 1990s saw the introduction of flexible docking systems from a number of groups, including the predecessor to Surflex-Dock, called Hammerhead [2], and others such as FlexX, Gold, and AutoDock [3–5], making use of a number of different approaches to scoring intermolecular interactions [6–8]. During the current decade, a number of methods have achieved fairly wide use, including Surflex-Dock [9–11] and other approaches, both academic and commercial, such as AutoDock, DOCK, Glide, Gold, FlexX, Fred, and SLIDE (for a review, see [12] or [13]).

In a theoretical sense, solution of the docking problem lies in correctly computing the combination of enthalpic and entropic effects that come from the formation and destruction of interactions among the protein, ligand, and solvent in the form of hydrogen bonds, Van der Waals interactions, formally charged interactions, and the entropy losses of the protein and ligand balanced against the entropy gains of the solvent. Direct methods exist to estimate ΔG_{bind} through the partition function, but these involve enumeration of all states of the system (bounded reasonably by energy) along with all corresponding energies [14]. An accurate picture of a protein/ligand interaction would involve an ensemble of the most probable protein and ligand conformations given an accurate calculation of the free energies attributable to each state. This is not feasible for many docking applications, given the speed requirements.

In a computational sense, due to the complexity requirements, the docking problem is typically formulated

as a search for a global optimum in a landscape that is defined by a scoring function, protein structure, ligand structure, and the degrees of freedom to be explored. The scoring function and search strategy combine to yield the solutions that a method will report. In nearly all high-throughput docking approaches, protein conformation is *not* among the degrees of freedom being searched. So, changes in the protein structure influence the *shape* of the energy landscape, not just the starting point of the search, and this also affects the solutions that will be reported. The energy landscape itself is usually characterized by a scoring function that is driven by inter-molecular energetics, treating intra-ligand energetics in a reduced fashion and largely ignoring the protein energetics. Generally, docking methods report a small number of poses, with evaluations tending to focus on either the accuracy of the geometric configuration of the top scoring pose or on some aspect of the score of the top pose (e.g. whether scores rank true ligands above non-ligands).

As the field has matured, use of shared benchmarking, especially by independent investigators, has become more common [12, 15–19]. This has revealed three key things. First, while a number of methods appear to produce similar performance in tests of geometric docking accuracy (roughly 60–80% success in producing correct top-ranked dockings of ligands to their *cognate* protein structure), the methods work *much less well* when making use of *non-cognate* protein structures (closer to 20–40% correct). Second, the methods are highly target dependent with respect to performance on pose prediction or screening enrichment. Third, there is no reliable correlation between predicted scores and binding affinities of ligands at the level required for guidance in lead optimization. A recent issue of the Journal of Computer-Aided Molecular Design was devoted to these issues, particularly these papers: [20–23].

The challenge of docking non-cognate ligands is illustrated in Fig. 1, where two ligands of PDE4b are shown [adenosine-5′-monophosphate (AMP) and 8-bromo-AMP]. The single atom change (hydrogen to bromine) results in a 180° flip of the heterocycle despite the fact that a more subtle shift could accommodate the additional steric bulk. Overall, the protein structure changes relatively little, with the largest shift being with a methionine sidechain in the active site. The flipped ligand is significantly easier to "predict" given the cognate protein structure for bromo-AMP than it is for the cognate protein structure for AMP itself, despite the relatively slight protein movements. When protein movements are slight, as in this case, an acceptable pose can often be identified among the top scoring set when docking against a non-cognate structure. However, the challenge lies in correctly discriminating the correct pose when the difference between top ranked (incorrect pose) and lower ranked (correct pose) is frequently <1 kcal/mol.

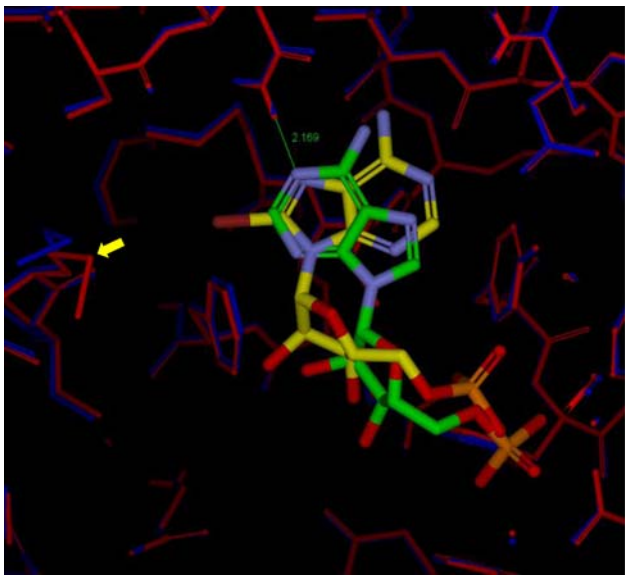

Fig. 1 Two PDE4b structures (1R09 and 1RoR) are shown superimposed. The former, shown in *red* is in complex with 8-bromo-AMP (*yellow* carbons), and the latter, shown in *blue*, is in complex with AMP (*green* carbons). The single atom change from hydrogen to bromine results in a complete flip of the adenosine, where a common hydrogen bond is made by different atoms on the heterocycle. The protein conformational shift is subtle, with the largest change being in the position of MET-431 (indicated with an *arrow*). However, docking into the *cognate* structure, the correct pose of 8-bromo-AMP is ranked much higher than when docking into the structure determined with AMP instead

The work reported here deviates from concentration on a single protein conformation, a single "best" predicted ligand conformation, and strict reliance on a scoring function that is dominated by inter-molecular effects. Protein conformational variation is considered on a large scale by using multiple protein structures for individual targets, and it is considered on a small scale by exploring local optimization of protein atomic coordinates in complex with a docked ligand. Instead of predicted ligand poses being considered separately, ligand *pose families* are considered, yielding ensembles of geometrically related poses whose ranks are determined in a probabilistic manner. In computing the scores of pose families, protein/ligand inter-molecular interactions are, of course, considered, but intra-molecular interactions, both non-covalent and covalent, for both the ligand and protein, are also considered.

Another, somewhat different, avenue involves the appropriate use of pre-existing knowledge in docking. The use of information about the bound pose of a cognate ligand in re-docking that ligand to the cognate protein can lead to serious problems of bias [23, 24]. However, in

practice, modelers seek to exploit their knowledge of well-studied ligands in making better predictions about new ligands, especially those that share structural features (e.g. a common P1 binding element for a serine protease inhibitor or a common hinge-binding moiety for a kinase inhibitor). In good hands, this can have a very positive impact on the performance of docking algorithms, but it can also lead to problems if overgeneralizations are enforced as hard constraints. The approach taken here makes use of small numbers of fragments of the cognate ligands from a small set of protein structures that are to be used to guide the docking of *new* ligands. The methods used are fully automatic and lead to no "contamination" of results, since the ligands to be used to evaluate performance are *never* used as information that affects the input to docking protocols. This approach leads to more efficient and deeper searching of binding modes that are related to those known to exist for ligands with common subfragments, but the constraints are not strict so alternative binding modes are explored as well.

Results are quantified for pose prediction accuracy in cross-docking, where ligands were docked into pharmaceutically relevant targets whose structural determination was done with *different* ligands. Eight targets, with a total of 211 test ligands, comprised the benchmark. Use of multiple protein structures per target with the standard Surflex-Dock scoring scheme yielded performance for top scoring poses of $\sim$50% correct ($\leq$2.0 Å rmsd), compared with roughly 25% correct using a single arbitrarily chosen protein structure. The level of performance seen with multi-structure docking is close to that of cognate docking on "hard" benchmarks (e.g. the 100 complex Vertex set [10, 18]). Through the use of post-docking protein pocket adaptation, pose family ensembles, and generalized scoring, examination of just two pose families per ligand yielded a mean success rate of 75% across the eight targets (single pose family performance averaged 64%). The level of performance obtained considering two pose families approaches what is observed on cognate docking (single top-scoring pose) with "clean" benchmarks (e.g. the 85 complex set of Hartshorn et al. [25]).

The approaches presented here are practical for use in lead optimization exercises. The docking protocol employing multiple protein structures takes just a few minutes per ligand. The rescoring protocol that performs protein pocket adaptation takes $\sim$30 s per pose per protein pocket when moving heavy atoms as well as protons. With five protein structures per target and ten poses per ligand, rescoring times were typically 30 min per ligand. While this is an expensive computation, use of multi-core, multi-node computing clusters means that sets of tens of ligands can be fully processed in less than an hour on widely available servers.

Methods and data

The present study makes use of two publicly available data sets to demonstrate improvements, both tangible and operational, in docking novel ligands to targets of pharmaceutical significance. Neither set was constructed for this study, rather being the work-product of third parties that were kind enough to share their data. Neither set was "cherry-picked" in any fashion. The following describes the molecular data sets, computational methods, detailed computational procedures, and quantification of performance.

Molecular data sets

Two data sets are used here to establish performance in geometric docking accuracy. The first, a cognate docking set, from Hartshorn et al. [25], contains 85 protein/ligand complexes. These were selected by the authors to represent a diverse, high-quality assortment in which questions about structure quality or uncertainty in ligand placement are at a minimum. The authors provided two alternative protein structures, one with protons optimized in the presence of the ligand with GoldScore, and one with ChemScore. For this work, the GoldScore variant (protein_opt_h_gs.mol2) was used (the other variant was not tested). Cognate ligands were provided as MDL mol files with all protons expected at physiological pH. These ligands were randomized (free torsions and alignments) and minimized to produce starting points for docking. This set will be referred to in what follows as the Astex85 set. This set was used primarily to establish an upper bound on how well docking can work in the case where the protein conformation is known to be maximally hospitable to the ligand to be docked. Note, however, that the proton optimization that was carried out was *not* done with the Surflex-Dock scoring function, so there is no particular bias in the proton coordinates that favored the minima that this scoring function prefers.

The second set, a cross-docking set, was provided by Jeffry Sutherland (personal communication). It consists of eight protein targets, each represented with up to ten different co-aligned structures from different protein/ligand complexes. For each target, the first five structures were used as input to molecular docking. A total of 211 ligands from *different* complexes were available for testing. Figure 2 shows all of the cognate ligands for PDE4b, CDK2, and ESR1 (above the line), and typical examples of non-cognate ligands used for testing below the line. The shaded moieties are geometrically equivalent in terms of their protein interactions. Figure 3 shows cognate ligands for thrombin (F2), MAPK14, and MMP8 (only four ligands are shown for two proteins due to space considerations).

The remaining two protein targets were PDE5a and MMP13, and structures are not shown in the interest of space. Results for PDE4b/5a and MMP8/13 are combined in what follows, since the total numbers of ligands for the target variants were small. This set will be referred to as the CrossDock211 set.

Protein structures for docking were prepared from the original PDB files and aligned to the structures that formed the original data set, due to a small number of errors in the original structures. These structures were optimized with their cognate ligands in order to eliminate large effects on computed internal energies of binding pockets that would otherwise result from differences between nominal optimal bond lengths, angles, etc. between those used for crystallographic structure solution and those used for scoring predicted protein/ligand complexes. Note that this *does not* result in contamination of prediction results, since the ligands used in the optimization process were *different* from those used in docking.

Ligand structures for docking were treated in two ways. For the 211 test ligands, docking was carried out using randomized test ligand conformations as well as using minimized versions of bound poses. The latter was done to simulate a typical modeling workflow, where the modeler builds ligands "in place" based upon a best guess as to the bound pose of a new ligand (including a sensible guess as

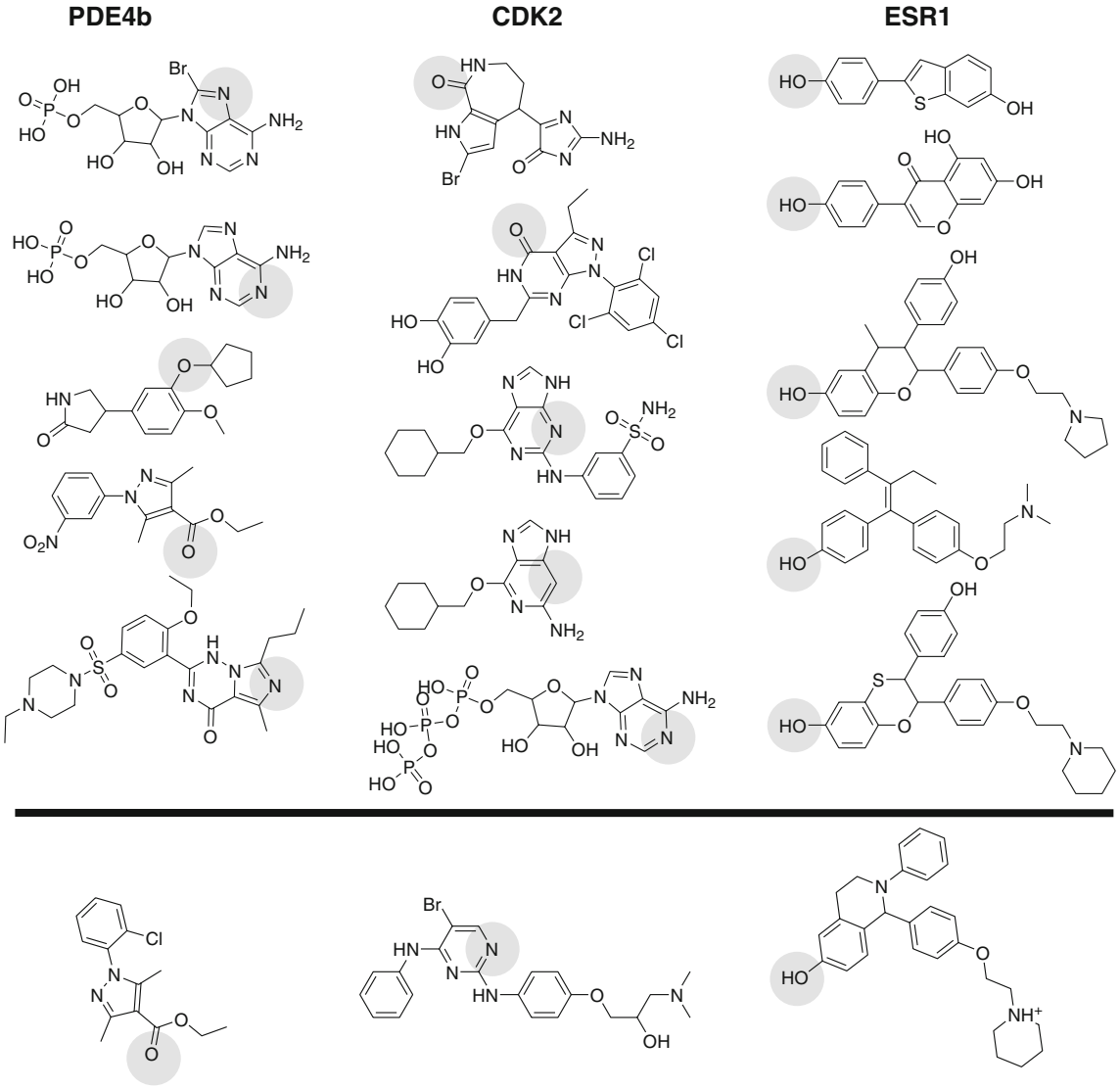

Fig. 2 Cognate ligands for PDE4b, CDK2, and ESR1, with example test ligands shown *below the line*. *Light shaded circles* highlight corresponding moieties on the ligands within each target (H-bond acceptor interacting with the sidechain amide of GLN-443 within PDE4b, H-bond acceptor interacting with amide proton of LEU-83 within CDK2, hydroxyl interacting with the carboxylate of GLU-353 within ESR1). Note that a single atom change (hydrogen to bromine) causes a 180° moiety flip among the top two ligands of PDE4b (the pyrimidine flips relative to the remainder of the ligand in order to roughly superimpose the *highlighted* nitrogens)

to tautomeric state and probable ring conformations). Results were slightly better using the latter scheme, as would be expected, but success rates in docking were not statistically significantly different either within a single target or over all targets (based on Fisher's exact test at $p = 0.05$ of successful docking with a threshold of 2.0 Å rmsd). In what follows, the results refer to the latter scheme unless otherwise noted.

All protein and ligand structures as well as preparation protocols are available for download (see http://www.jainlab.org for details).

Computational methods

The core computational methods within Surflex-Dock have been reported in previous papers and will be described only briefly here. Those methods that represent modifications and enhancements will be presented in detail.

Scoring function and search strategy

Surflex-Dock employs an empirically derived scoring function, where the parameters of the function are based on protein-ligand complexes of known affinities and structures. The function may also be tuned by using information from non-binding ligands or hard docking failures (see [11, 26, 27] for extensive details on the Surflex-Dock scoring function as well as a review of its relationship to other approaches). Conceptually, the scoring function, as with the entire family of empirical scoring functions, borrows heavily from the approach of Bohm [6], with terms for hydrophobic contact, polar interactions, and entropic fixation costs for loss of torsional, translational, and rotational degrees of freedom. However, the Surflex-Dock scoring function makes a significant departure from other approaches in two important respects. First, the function is composed of a sum of *non-linear* terms and it is continuous and first-order piecewise differentiable. Second, the parameter estimation regime for the function takes direct account of the problem of ligand pose variation. Very small changes in ligand pose can yield large differences in the nominal value of a scoring function. Rather than taking the precise pose from a crystal structure, the approach is to find the nearest local optimum and define the score at that optimum as the score for the ligand. This follows the approach developed for Compass, which established the conceptual framework for this approach, termed *multiple instance learning* within the computational machine learning field [28, 29]. For a more detailed discussion of the Surflex-Dock scoring function, please refer to the specific reports of the derivation and refinement of the function [7, 11, 26].

A detailed account of the Surflex-Dock search algorithm can be found in the original paper [9], and additional refinements were described in a more recent publication [10]. The method employs an idealized active site ligand (called a protomol) as a target to generate putative poses of molecules or molecular fragments. The protomols utilize CH_4, C=O, and N–H molecular fragments. The molecular fragments are tessellated in the protein active site and optimized based on the scoring function. High scoring fragments are retained, with redundant fragments being eliminated. The protomol is intended to mimic the ideal interactions made by a perfect ligand to the protein active site that will be the subject of docking. Surflex-Dock utilizes a molecular-similarity based alignment engine to generate putative alignments of fragments of an input ligand to the protomol. Poses of the molecular fragments that tend to maximize similarity to the protomol are used as input to the scoring function and are subject to thresholds on protein interpenetration followed by local optimization. The partially optimized poses of the fragments form the basis for further elaboration of the optimal pose of the full input ligand. The procedure identifies high scoring fragments that have compatible geometries to allow for merging in order to construct a high scoring pose of the full input ligand. The whole molecules that result are pruned based on docking score and are subjected to further gradient-based score optimization. The procedure returns a fixed number of top scoring poses.

Recent improvements to this basic procedure include implementation of a covalent force-field, which supports all-atom Cartesian ligand optimization, either before or after docking, as well as a general approach to ring flexibility. Screening performance can be dramatically improved by making use of docking protocols that employ these methods [10]. The other recently reported improvement with specific relevance to the work reported here is a procedure for making use of molecular fragments of known binding geometry to help guide docking. Frequently, one is exploring a chemical structural space of analogs of well-studied series of compounds, as is modeled in the cross-docking data set under consideration. In these cases, it is reasonable to posit that a particular substructure has an especially favorable interaction within an active site (as with, for example, metal chelation moieties), making direct use of that knowledge to focus the search offers advantages in terms of workflow, speed, and direct comparison of different analogs. Using this procedure, one can specify a *collection* of placed molecular fragments. In cases, where a ligand to be docked contains a particular substructural fragment, the known geometry of that fragment is used to explore the space of docked poses in which the matching part of the ligand is congruent with the placed fragment.

Importantly for the work reported here, the fragment-based docking approach is not used *in place of* the standard unbiased docking protocol, but in addition to it. So, placed

fragments ensure that known binding geometries of particular moieties are explored, but alternative dockings that score higher will be reported as well. From the example in Fig. 1, knowledge of the binding geometry of the AMP heterocycle does not prevent identifying the correct binding mode of the 8-bromo derivative. The primary focus of the current study is on improving performance in cross-docking geometric accuracy. So, clever choice of *which* fragments to use in the docking procedure could yield very significant effects on performance. Consequently, an automated computational procedure was implemented that made use of *only* the cognate ligands of the five protein structures for each target. Figure 4 shows the automatically chosen fragments for PDE4b that interact with GLN-443. An additional two fragments (not shown) represent other, less central interactions.

Protein conformational variability

Several groups have approached the problem of protein flexibility in docking, with a number of notable successes. McCammon's group introduced the relaxed complex

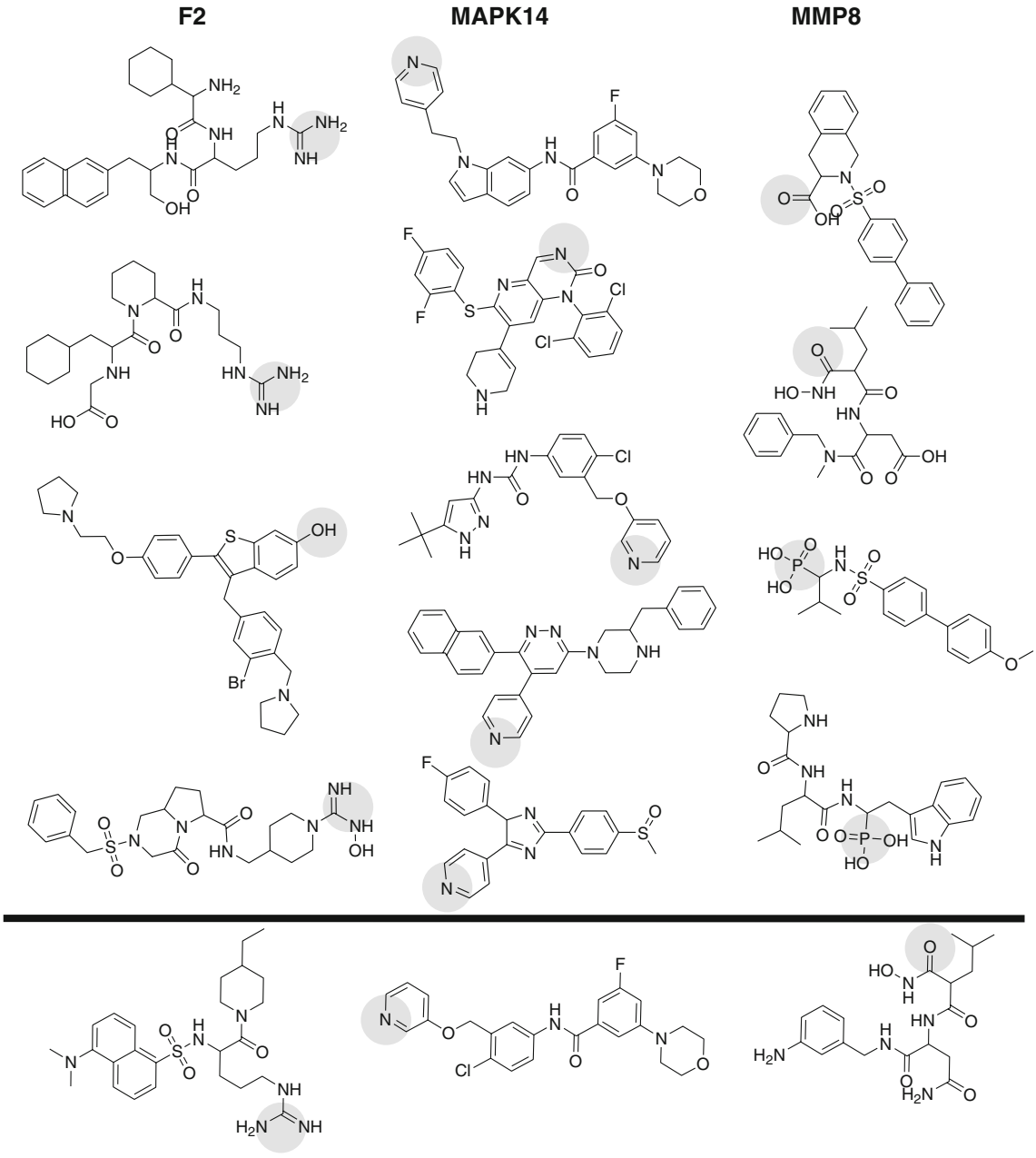

Fig. 3 Cognate ligands for F2 (thrombin), MAPK14, and MMP8, with example test ligands shown *below the line*. The *highlighted* moieties indicate corresponding functionality for each target (P1 element for F2, H-bond acceptor for the amide proton of MET-109 in MAPK14, and metal chelation element for MMP8)

method, which made use of multiple protein structures sampled from molecular dynamics, and both the initial work and subsequent refinements has shown the utility of using multiple structures coupled with sophisticated scoring schemes [30, 31]. Osterberg et al. [32] showed how explicit considerations of residue movement (and structural waters) can substantially increase docking accuracy for AutoDock in HIV protease. The approach made use of a combined grid of interaction energies instead of a single one. Kairys and Gilson [33] reported on an extension to the Mining Minima method where, on a number of protein targets, mobility in hydroxyls and sidechains improved docking accuracy. Cavasotto and Abagyan [34] extended the ICM approach to allow for multiple discrete protein conformations (which included backbone changes), and they showed improvements both in screening utility and cross-docking accuracy. Shoichet's group established that accounting for variations in energy among different protein conformations can lead to significant improvements in screening utility [35]. More recently, the developers of Glide have made progress in sampling of sidechains and limited backbone movement *within* a docking process that employs iterations of ligand sampling and protein sampling [36], but the process takes *hours* for a single docking of a single ligand.

The approach here was informed by this earlier work. However, rather than relying on computational methods to sample large motions of proteins, multiple experimentally determined protein structures are used. Large motions, such as those encountered with agonist versus antagonist bound forms of nuclear hormone receptors are extremely challenging to predict accurately enough for operational use in lead optimization guided by docking. In practical situations, in lead optimization exercises that are being guided by in-house crystallography, the larger motions are likely to be captured by experimental structure determinations. Smaller motions, including both sidechain and backbone atomic movement, are explored in Cartesian space with a blended scoring function that includes both the non-covalent intermolecular forces as well as the covalent *and* non-covalent intramolecular forces for both the ligand and the protein. This approach takes a small number of initial protein conformation samples (five in this work) and makes use of local optimization after docking in order to gain the effect of finer sampling of protein conformational space.

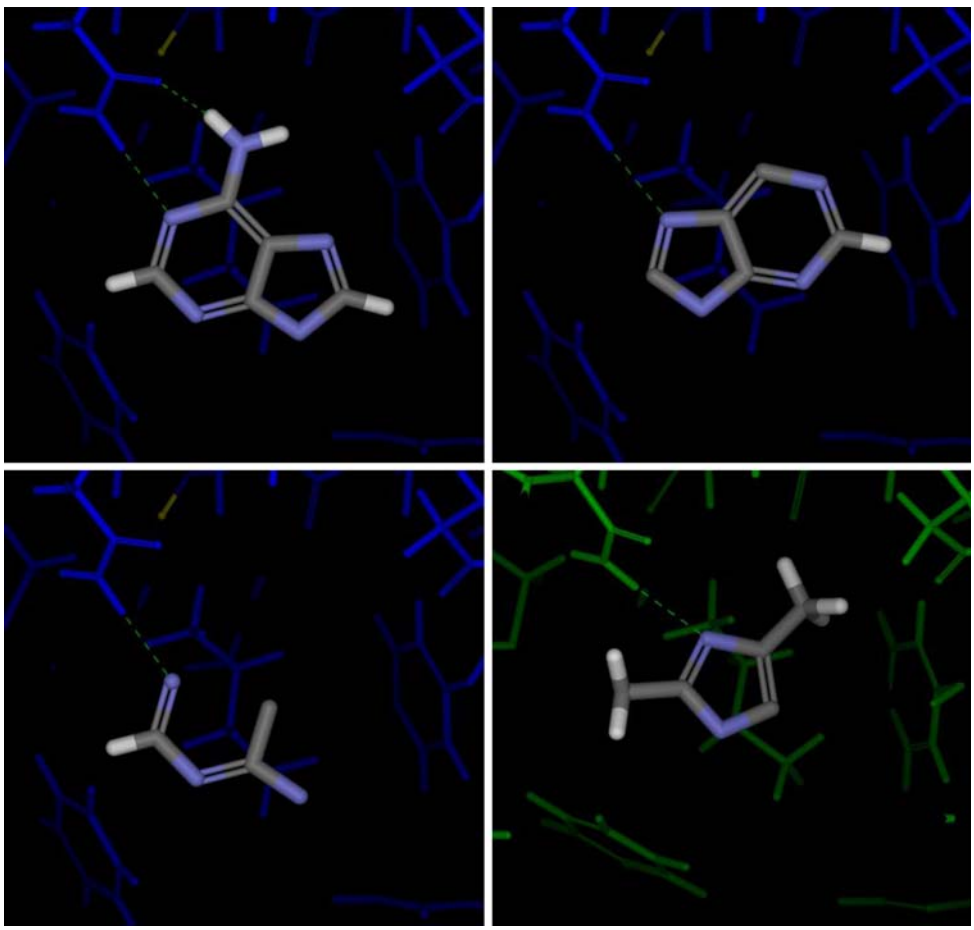

Fig. 4 Automatically chosen fragments of the cognate ligands for PDE4b. Each makes a critical interaction with GLN-443. The fragment on the lower right is able to make a hydrogen bond with its cognate protein conformation (1XOT, shown in *green*), whereas the other three are able to make interactions with the same protein conformation (1RO9, shown in *blue*)

For a single ligand docked to N protein conformations, with M poses returned, rescoring each of the poses is performed against all N protein conformations, optionally with K small random perturbations in order to generate a fine-grained sampling of the scores to be expected. For the results reported here, N was five, M was ten, and K was zero (no random perturbations) in order to keep computational costs low, but this still resulted in 50 alternative protein/ligand complex configurations per docking. Better results are possible with increased sampling (e.g. M of 20 and K of 4), which results in 500 configurations per docked ligand. One of the features of such sampling is that rare configurations can occur, which have nominally favorable energetics, but which lie within a very tiny slice of configurational space and therefore are not the most probable biologically important pose. To address this sampling issue, and to produce an improved workflow where a very small number of solutions must be considered, rather than reporting results on single configurations, *pose families* are constructed that surround significantly different central poses. These families are scored using a Boltzmann weighted probability scheme, with pose families with high probability ranked above those with lower probabilities.

These procedures are detailed as follows.

Multi-structure docking: The generalization from single protein conformation docking to multiple conformations was straightforward, simply iterating the docking process that has been described extensively in prior reports [9, 10]. The implementation allows specification of a set of protein structures, each with one or more protomols, in a single file. Each independent docking shares a final pose set of fixed size (default of 20), which contains the best poses based on intermolecular non-covalent scores over all protein conformations. In this process, no movement of protein atomic coordinates occurs. All of the options that control standard docking (e.g. pre-docking ligand minimization, post-docking all-atom ligand optimization, dynamic ring search, etc.) are available.

Note, however, that there are opportunities for additional efficiencies that will be pursued in future refinements. In particular, since the protein structures are aligned in a common coordinate frame, the process of ligand pose generation need not proceed independently for each of the individual protein conformations. Instead, generation of putative alignments could take place once, with the alignments being scored within each pocket variant separately. The focus of the work reported here has been to establish the feasibility of an operationally practical workflow rather than an optimal one in terms of computational efficiency, so such refinements remain as future work.

Protein pocket adaptation: The mechanical aspects of protein pocket adaptation were implemented previously, in order to study the bias effects of protein coordinate optimization on *cognate* docking [10, 23]. The process is straightforward. For a particular ligand pose within a particular initial protein conformation, the protein atoms near the ligand (those whose van der Waals surface distances are <4.0 Å) are identified and marked. If the selected protocol calls for moving protein protons only, then heavy atoms are unmarked. In all cases, protein atoms that chelate metal ions are unmarked (as are the metals themselves). A scoring function is instantiated that includes three terms: (1) the inter-molecular non-covalent components of the Surflex-Dock scoring function; (2) the intra-molecular non-covalent terms of the Surflex-Dock scoring function (for both the ligand and the protein); and (3) the intra-molecular *covalent* terms for both the ligand and the protein. The covalent terms for the protein include all bond length, bond angle, and torsional terms where at least one atom of the protein is marked. The total complex score (computed as kcal/mol) is minimized. The resulting score is reported in several ways, including the total score, the separate components, an estimate of ligand strain, and a scaled complex score (called "CScale") that normalizes the protein score components so that ligand poses that contact different numbers of protein atoms are more directly comparable.

The implementation of the functions includes analytical computation of gradients, and the optimization itself is carried out using a modified quasi-Newton scheme [10]. During the optimization process, the gradients for the protein atoms that are unmarked (and therefore not supposed to move) are zeroed. All atoms to be optimized are moved simultaneously in the procedure (an earlier implementation iterated protein movement with ligand movement). Selectable parameters control the weighting of the protein covalent force-field (here set to 0.6) and the ligand covalent force-field (here set to 1.0) and whether or not non-covalent intra-molecular interactions should be included (here these were included). Systematic optimization of parameter choices was not carried out; instead, a small number of complexes from the previously studied Vertex docking set were used to identify acceptable parameters for application to the data in this study [10, 18]. Typical run-times for optimization of single complexes (all pocket atoms) on a single-processor were 30 s of wall-clock time on standard Intel-based hardware running on Linux systems (e.g. 2.0 GHz Core 2 T7200).

There are a number of potential efficiencies to be pursued to improve over the current implementation of serial optimization of multiple ligand poses against multiple protein conformations. Some are purely technical, having to do with local optimization approaches that scale more efficiently than the current one in the number of parameters under optimization. Others will involve extensive precomputation of subtle variations in protein conformations

and storage and use of intermediate results for use in subsequent optimization steps. Even without further improvements, the procedures are operationally feasible, requiring roughly 30 min per ligand on a single processor in the protocol used for the results presented here.

Pose family clustering: Input to the pose family procedure is a set of ligand poses along with a set of scores. Here, the scores were taken as the standard Surflex-Dock scores (converted to kcal/mol) for pose family computations on original dockings and the CScale scores mentioned earlier for poses resulting from pocket optimization (with either protons only or all protein pocket atoms). For each pose of the Q total number of poses, a Q-dimensional binary vector is computed, with values set to 1 for those poses that were similar (<1.5 Å rmsd) to the pose in question. Each pose is also assigned a probability based on its Boltzmann-weighted share of the total from all poses. Each pose, along with its marked neighbors, *may* form a pose family for ranking and output, but the pose families are produced from most probable (total probability over all poses within the pose family) to least, and less probable pose families that are similar to more probable ones are skipped. The similarity threshold is user settable and is expressed as a Tanimoto similarity between the binary pose family vectors. For this work, pose families had to be nearly non-overlapping (Tanimoto <0.05) in order to survive the process. Also, in order to "thin" the number of poses produced per pose family and focus attention upon those poses that had meaningful contributions, the contribution of a pose to the overall docked ensemble had to be greater than an individual probability of 10^{-6} in order to be shown in the output structure file comprising the pose family.

This computation did not add appreciably to the total times for ligand processing. The net result of these procedures was, for each ligand, three ranked sets of pose families, with one from the initial docking, and one for each of the two methods of rescoring with pocket adaptation (protons only or all atoms). In what follows, only the top-ranked pose families from each scoring method were used.

Computational procedures

Details of computational procedures in studies, such as this can have a remarkable impact on results, both with respect to the actual performance of algorithms but also as to the comparability of different methods that have been run on nominally the same benchmarks. The publicly available data archive associated with this paper contains all protein, ligand, and protomol structures as well as example scripts for the primary experiments described. The following summarizes the procedures used at a level of detail intended to give a clear picture of the key choices made for the current study.

Astex85 set preparation: For the Astex85 set, proteins were used unmodified, with cognate ligands being subjected to torsional randomization followed by minimization prior to docking. Protomols were generated using default procedures, as described previously [10].

CrossDock211 set preparation: For the CrossDock211 set, protein preparation for docking relied upon an automated procedure for generating SYBYL mol2 files from original PDB files, resulting in protonated proteins and ligands, with tautomeric states being enumerated and chosen to yield complementary bound states. Ligand bond orders were automatically assigned and were reviewed manually to correct the small number of cases where the automatic assignment was incorrect. Proteins and ligands were transformed to a common alignment based on the structures from the original data set. Protein active sites were trimmed to include residues within 15 Å of the cognate ligand. The resulting complexes were then optimized in two different ways, one allowing for protein pocket adaptation of protons only and the other allowing for all pocket atoms to move (using the Surflex-Dock "popt" command).

For each protein structure prepared as described, protomols were generated using default procedures, with the union of cognate ligand structures for all five protein conformations used to identify the scope of the active sites. Generation of molecular fragments was done automatically based solely on the structures of the five cognate ligands for each target using the Surflex-Dock "fragmentize" command. Selection of which fragments to use to guide docking was also fully automatic, operating on a collated set of fragments from all cognate ligands and on a collated set of the ligands themselves (using the "choose_frags" command). The 211 non-cognate test ligands were used as provided in the original data set, followed by automatic protonation/minimization, optionally including torsion randomization prior to minimization.

Docking procedures: Baseline results for both the Astex85 and CrossDock211 sets were generated using default geometric docking parameters with a *single* structure per protein (e.g. sf-dock.exe -pgeom dock_list test.mol2 p1-protomol2.mol2 protein.mol2 log). For the Astex85 set, this was a *cognate* docking test. For the CrossDock211 set, both cognate and *non-cognate* baseline results were generated. The non-cognate results employed the 211 novel ligands, and the cognate test employed the protein/ligand complexes used as targets for the cross-docking experiments (five structures for each of eight targets). The primary results of the study involved multi-structure docking on the CrossDock211 set using the cognate molecular fragments to help guide search,

followed by protein pocket optimization for each docked pose, and generation of pose families. The procedure that made use of heavy atom pocket adaptation was as follows:

1. sf-dock.exe -div_rms 0.25 -fmatch cdk2/train-ligands/chosenfrags.mol2 mdock_list cdk2/test.mol2 Targets-cdk2 cdk2/log
   This command performs a multi-structure docking, with a guarantee that no output poses will be <0.25 Å rmsd from any other, using the placed fragment specified to guide docking. The pathnames to the five protein structure files along with their corresponding protomols is in Targets-cdk2.
2. sf-dock.exe -ntweak 0 -maxposes 10 +pflex +hprot -pcov 0.6 +self_score rescore_multi cdk2/log Targets-cdk2 rescoreheavy
   This command rescores a multi-structure docking run, using no random perturbations of the final dockings, considering a maximum of ten poses per ligand, with protein pocket flexibility including heavy atoms (covalent force-field weight of 0.6), and where intra-molecular interactions count in the scores along with the inter-molecular interactions.
3. sf-dock.exe posefam cdk2/log-rescoreheavy
   This command generates pose families for all ligands from the docking run, based on the scores in the log file along with the associated poses in an archive prefixed with the log file name.

The result of the sequence of operations was a set of pose families for each test ligand (e.g. log-rescoreheavy-ligand-1-fam-*.mol2). In the results that follow, "baseline" performance refers to the multi-structure docking with no pocket adaptation or rescoring, and two forms of rescoring with pocket optimization refer to moving heavy atoms or just protons (analogous to the above procedure but with "−hprot" instead of "+hprot").

Results and discussion

In multiple reports, a group of docking methods (Glide, GOLD, and Surflex-Dock) performed close to equivalently with respect to docking accuracy. The absolute performance varied based on the benchmark. Percentage of top-scoring correct poses (≤2.0 Å rmsd) in the cognate docking problem ranged from 50 to 60% in a 100 complex benchmark from Vertex [10, 18]. The percentage of correct poses within the top 20 returned (but not necessarily top-ranked) ranged from 75 to 85%. On an independently run benchmark of 100 complexes from Rognan's group comparing eight docking methods [19], the comparable numbers for the three methods were about 55% and 75–80%. On a benchmark constructed with very careful attention to quality of crystal structures (resolution, density covering the ligands, etc.) from Hartshorn et al. [25], GOLD performed at 71–87% correct for top scoring correct poses, depending upon the precise conditions (binding site definitions, initial ligand geometry, search depth, etc.). In the much more relevant cross-docking situation, performance for all methods is quite a bit lower, but with the same methods performing well. Warren et al. [12] studied eight targets using several docking methods, with additional methods tested subsequent to the original publication [37]. Comparing the average rank-order of performance across the eight targets, among Dock4, Dockit, FRED, FlexX, Flo, GOLD, Glide, Ligfit, MOE, Surflex-Dock, the top three (in order) for top-ranked pose were Surflex-Dock, GOLD, and Glide and for best pose were GOLD, Surflex-Dock, and Flo (with Glide coming in fourth). However, the *absolute performance* was significantly worse.

So, in cases where we can guarantee that a protein structure is near-optimal for the *particular* ligand being docked (e.g. as in the Hartshorn study), we observe very good performance: nearly 80% correct for top-scoring poses. As the quality of structures becomes more variable, even in the cognate docking case, the performance is reduced to about 55% for multiple methods (e.g. on the Vertex data set). As we move to the operationally important cross-docking case, that of docking a novel ligand into a protein whose structure was determined with a *different* ligand, we see a further significant reduction in prediction accuracy. Figure 5 illustrates this point on the Astex85 docking set and the CrossDock211 set. In the cognate docking case, without any optimization of docking protocol, Surflex-Dock achieved 76% correct for top scoring poses at the 2.0 Å threshold, with over 95% of the dockings having a correct solution within the top 20 poses returned. However, in the cross-docking case, top scoring pose accuracy decreased to 25% and best pose success dropped to 60%.

Note, however, that the comparison of cross-docking on the CrossDock211 set to cognate docking on the Astex85 set represents a hardest-case to easiest-case comparison, since the Astex85 set was cleanly curated to include particularly high-quality structures by multiple criteria, apart from being a cognate docking test. Performance of Surflex-Dock was evaluated on the *cognate* protein/ligand structures from the CrossDock211 set as well. Success for top-scoring pose was 65% and for best pose of top 20 was 90%, which was lower than that observed with the Astex85 cognate-docking (76% and 95%, respectively), but not statistically significantly so (by Fisher's exact test at $p = 0.05$). In a similar comparison, Verdonk et al. [38], considered cognate docking on their Astex85 set with cross-docking of novel ligands into 65 of the 85 protein structures. They observed cognate docking performance

(top scoring poses ≤2.0 Å rmsd) of 80% for the 65 cognate cases, with a reduction to 61% for the cross-docking performance. This reduction in performance, while significant, was much less than observed here for Surflex-Dock on the CrossDock211 set. Apart from their set containing different targets and different ligands, they also included only those structures that contained the same set of binding site atoms present in the cognate structures and where the novel ligands were bound to protein forms that closely matched the reference structures in terms of protonation states, tautomers, and side-chain flips. Sutherland et al. [39] published cross-docking results for CDocker and Fred on the set used here, with success rates for top-scoring pose prediction ranging from 16 to 26%, paralleling what was observed here for *single-structure* cross-docking. Both groups considered the improvements possible by making use of using multiple structures, as will be done here in what follows.

As discussed above, there are marked differences in docking accuracy as we vary the degree to which we can expect the protein conformation to be "correct" for the purpose of accurately identifying the binding mode of a ligand. Proteins vary in their degrees of binding pocket flexibility, and some protein conformations can provide an inhospitable geometry for docking particular ligands. In the operational application of docking, we are *never* docking a ligand into the structure of a protein whose geometry is known, a priori, to be optimally complementary for the bound ligand. Figure 6 shows the degree of conformational variation for PDE4b and CDK2 among five different experimentally determined complex structures. PDE4b is comparatively rigid, but as we saw in Fig. 1, even small motions can influence docking preferences. CDK2 is clearly much more flexible, creating a more significant challenge in the cross docking scenario. Estrogen receptor (not shown) forms a middle ground, with relatively little variation among agonist-bound forms or antagonist-bound forms, but the differences between the agonist and antagonist forms are large.

Effects of multiple structures and fragment knowledge

Figure 7 shows the effect of moving from a single protein structure to five per target and of making use of placed fragments from the cognate ligands of the five protein structures to help guide docking. There is a statistically significant improvement through the use of multiple protein structures under the same docking protocol as used for

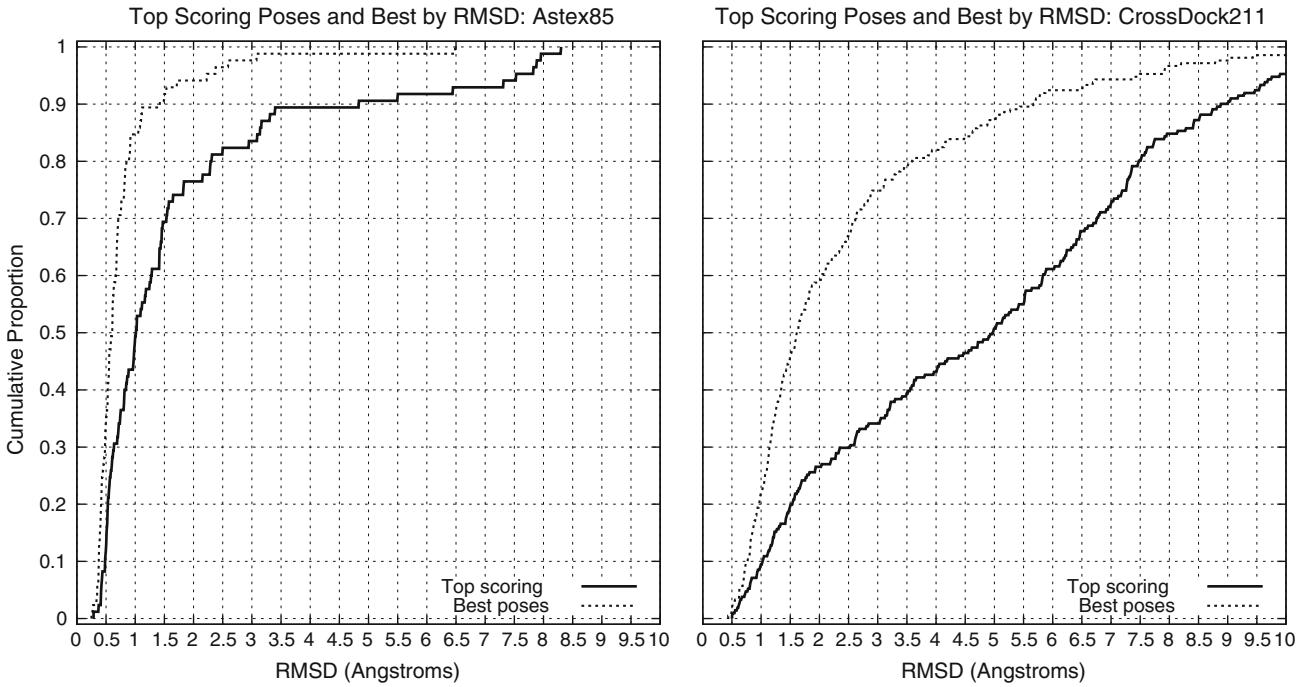

Fig. 5 Performance of Surflex-Dock using default geometric docking parameters on the Astex Diverse set of 85 cognate protein/ligand complexes (*left plot*) and on the cross docking set of 211 novel ligands docked to eight different protein targets (*right plot*). Overall performance in cognate-docking for top scoring poses was 76% success at a 2 Å rmsd threshold (cumulative histogram shown with *solid line*). The best pose of the top 20 was within 2 Å 94% of the time (*dashed line*). This level of performance is statistically indistinguishable from that of GOLD from the original paper. For cross-docking, the comparable performance levels were 25 and 60% for top-scoring and best pose, respectively. Among the cases where a good pose existed among the top scoring, the success rate for cognate-docking was 81%, but for cross-docking was 42%, highlighting the difficulty in ranking among poses under the latter real-world conditions

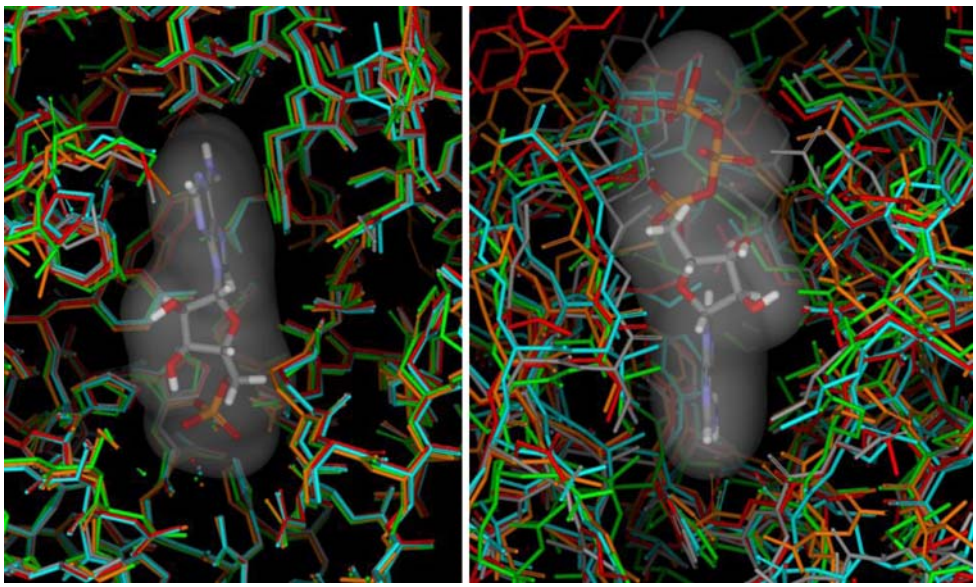

Fig. 6 Protein flexibility is significantly different among the targets. At *left*, the five conformations of PDE4b are shown along with a single ligand. At *right*, five conformations of CDK2 are shown, also with a ligand. PDE4b exhibits very little movement overall and has relatively little backbone variation. CDK2 contrasts by exhibiting movement in all atoms

single structures (green curves vs. red curves in the plots). Success rates at the 2 Å rmsd threshold improved to 45% from 27% for top scoring pose and to 82% from 60% for best pose by rmsd of the top twenty returned. Note, however, that median docking times increased by fivefold, since the procedure itself is essentially a sequential docking to each structure, with minimal additional efficiencies.

By making use of sub-fragments whose bound geometry is known, the process of docking is faster, and the space of solutions that share common features with known ligands is searched at a greater relative depth. Using this approach, top scoring pose success (at 2.0 Å rmsd) increased to 50% from 45%. This is not significant in a statistical sense at a single threshold (e.g. by using a test of difference of proportions), but the overall shift in the distribution of rmsd values is marginally significant. Interestingly, the distributions of rmsd values for best poses was essentially unchanged under the two conditions. The primary effect of the use of fragment knowledge was deeper search within the space of a priori favorable poses, which resulted in the slight improvement in top scoring pose identification.

Docking using the constraint of multiple fragments is relatively fast, and it eliminates the need for docking from multiple initial ligand conformations (which is done in the standard docking protocol for geometric accuracy). In the multi-structure protocol yielding the best performance in Fig. 7 (the blue curves), the overall docking speed was just 1.7-fold longer than the single-protein method. The median time to dock each ligand was just 4 min, with ligand flexibility ranging widely, but with typical ligands having roughly seven rotatable bonds.

The performance levels shown here represent a lower-bound in the sense that, while the docking protocol was designed to mimic that of an actual modeler making use of knowledge of multiple structures and well-understood interactions, the choice of protein structures was arbitrary, and the choice of fragment hints was made using no deep knowledge of the systems. For example, in the case of thrombin, a reasonable modeler would ensure that all common P1 binding elements would be represented among the fragment hints to be used by a docking system. Here, however, neither the very common amidine nor the more recent non-basic chlorophenyl P1 pocket binding elements were among the fragments used in docking, whereas they were very common in the test ligands (e.g. the ligands of 1KTS and 1WAY). In addition, while a number of ligands of thrombin were present that require chelation of a metal ion such as Zn^{2+} (e.g. the test ligand from 1C1W), none of the five example protein structures contained the required chelation moiety. In a real-world modeling exercise, when designing around a common binding element such as the P1 element in thrombin or any known metal chelation moiety, one would include preferred binding modes for those ligand components.

Verdonk et al. [38] showed a modest increase in top-scoring pose prediction (from 61 to 67%) in a multiple structure approach on their highly curated cross-docking data set. Sutherland et al. [39] showed a more substantial improvement on the data set used here, from 16% to 26% success for single-structure cross-docking to 36–46% success for multiple structures depending on the method of arbitration used to choose among the multiple dockings. Results for Surflex-Dock were of similar magnitude in terms of relative improvement (from 27% to 50%). Note that the protocol used here with Surflex-Dock made use just five alternate protein conformations per target, chosen a priori, whereas that used by Sutherland et al. used an all-by-all cross-docking.

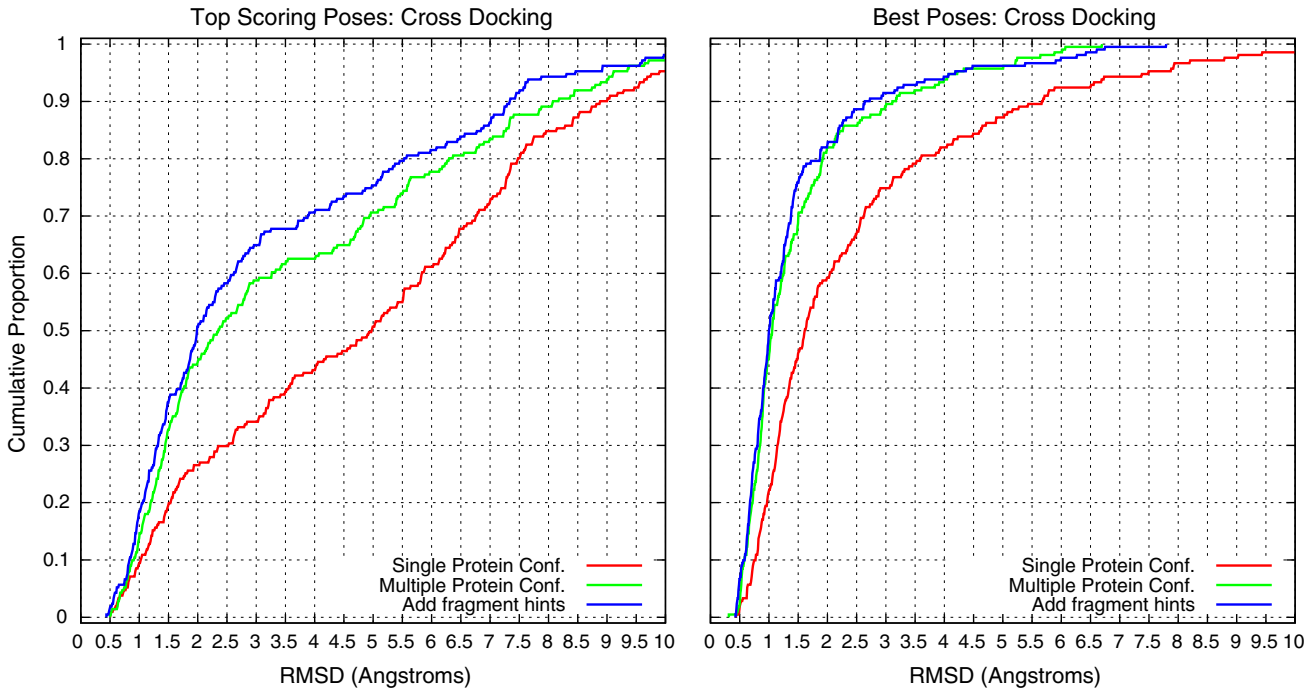

Fig. 7 The *left plot* shows performance of Surflex-Dock in cross docking for the top scoring returned pose for each of 211 non-cognate test ligands across eight different targets. The *right plot* shows performance considering the best pose returned among the top 20. Performance shows a substantial improvement resulting from the use of multiple protein conformations over a single protein conformation (*green* and *blue* curves, respectively). Making use of fragments from the structures of the cognate ligands of the protein conformations leads to deeper exploration of the likely to be correct pose space, which improves performance further (*blue* curve). The multi-structure docking protocol with fragment hints is nearly as fast as the standard single-protein Surflex-Dock protocol with geometric search parameters

Effects of protein pocket adaptation and pose families

The results thus far have deviated from most widely used docking methods and protocols by using multiple protein structures, but these structures have been treated as completely fixed. Further, top scoring poses have been treated as *the* singular solution to the docking computation. It is well understood that protein/ligand complexes are not accurately portrayed as the singular snapshot one often sees in a high-resolution crystal structure. Even in the case where a single joint configuration dominates others by having substantially lower free energy than significantly different configurations, the complex exists as an ensemble of configurations over short time scales where the coordinate changes may be small but are nonetheless real.

Figure 8 shows the docking of the CDK2 ligand from 1HO8 into the five protein conformations. At left is the ligand and protomol for 1OIU (one of the five structures used), with a particular subfragment of the cognate ligand shown in thicker sticks. That particular fragment was responsible for helping to guide the docking of the test ligand, which is shown in the middle panel in two poses (atom color), along with the fragment (blue), and the two alternative bound poses of the ligand from the experimentally determined structure (green). In this depiction, the effects of pocket adaptation are shown (red protein structure at right) along with the effects of identifying pose families. For the results of docking without pocket adaptation, the top scoring pose was 2.0 Å rmsd distant from the further of the two experimentally determined ligand poses. Pose families derived from the initial docking failed to group the two alternate solutions together. The closest poses to the correct pair of experimental ones were too far apart given the original protein coordinates. Rescoring the final pose set with full atomic adaptation within the binding pocket yielded significant movement, especially in the position of a key carboxylate. Generation of pose families from the rescored pockets identified a *single* pose family as being highly probable, with contributions from modifications of three parent protein structures. This top pose family contained conformations $<$1.0 rmsd from each of the experimentally determined alternatives.

Figure 9 shows all of the poses from the top scoring pose family resulting from pocket adaptation. They exhibit reasonable movement in light of the known variation in the "tail" of the ligand in question. Note, however, that the protocol using full atomic movement identified the correct

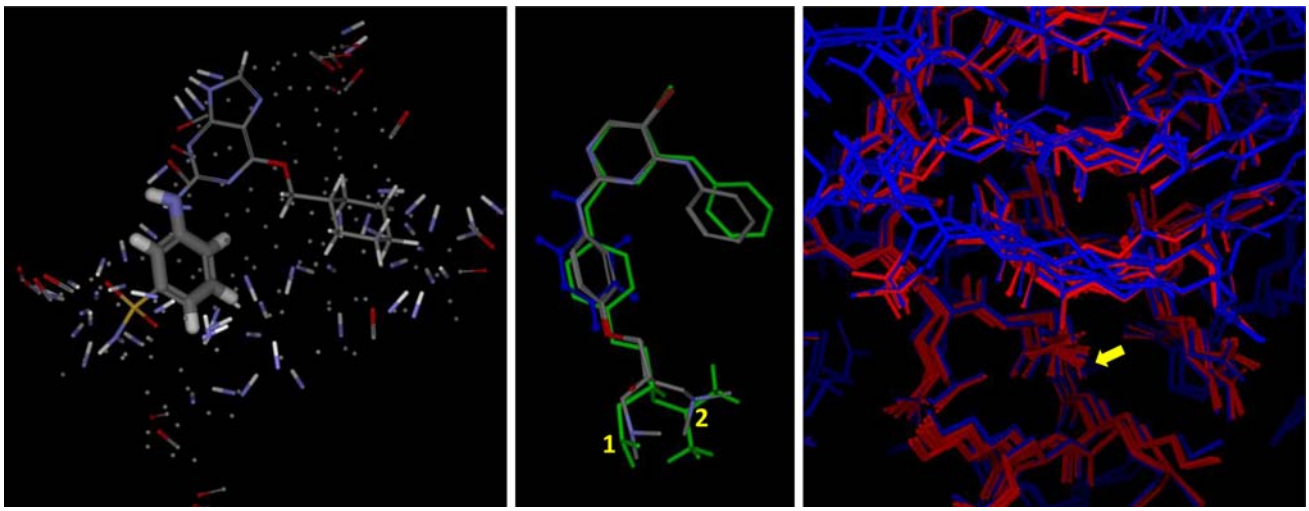

Fig. 8 Cross docking of the ligand from 1H08 into CDK2. At *left*, the protomol, cognate ligand, and a subfragment are shown. The protons from the steric protomol probes have been hidden, and the subfragment is shown with fat sticks. In the middle, two conformations from the top scoring pose family (resulting from heavy-atom pocket refinement) are shown along with the crystallographic alternatives (*green*, with the alternate amine positions numbered *1* and *2*) and the fragment that helped guide the docking (*blue*). At *right*, the three protein structures that contributed to the final pose family are shown in *blue* (1DM2, 1H0W, and 1OIU), with the refinements due to post-docking optimization shown in *red*. There are significant movements in the protein that allow the recognition of this pose family as being optimal, particularly near the carboxylate by the *arrow*. Pocket refinement with protons yields an incorrect pose family, and the pose family without pocket refinement does not span both solutions

family, but the protocol that was restricted to protons only identified the incorrect family as most probable (the second most probable contained one of the two correct alternatives). The plot at right in Fig. 9 shows the improvement in docking accuracy obtained by making use of top-scoring pose families instead of single top-scoring poses. Baseline performance (no pose families, just the top scoring pose) is shown in red, with some improvement seen in computing pose families without any pocket adaptation (purple line) that is due mostly to the difference in reporting method. For pose families, the minimum rmsd to experimental is computed, so there is a bias toward nominally better results, especially at the lower end of the curve.

All three methods of pose family generation (no rescoring, rescoring with proton movement in the protein pocket, and rescoring with all atom movement) yielded very similar performance at the 2.0 Å threshold: ~55%. This level of performance approaches that seen in cognate docking on "hard" cognate docking benchmarks (see earlier discussion), and the characterization of results resembles a sensible physical interpretation of protein/ligand binding.

Pose family agreement

As illustrated by the example from Figs. 8 and 9, the different scoring methods can yield different results, but their overall performance is close to equivalent. Since the scoring methods are computing only partially related terms, orthogonal agreement might suggest higher confidence. Figure 10 shows the relationship between top scoring pose family agreement among the three methods and prediction accuracy. Pose family agreement was calculated between each of the pocket adaptation top families and the top family from the original docking, with the mean deviation characterizing overall agreement. There is a striking relationship between nominal agreement and the accuracy of the top scoring baseline pose family. In over half of the 211 test cases, the three methods had highly similar top scoring pose families, and within that subset, the proportion of correct predictions was 80%. In an operational sense, this is a helpful feature, since it allows for confidence to be based upon the stability of the original top scoring pose family to protein pocket adaptation. This level of success is comparable to that seen with carefully selected and curated cognate docking sets (e.g. as in Fig. 5, with the Astex85 set).

In the remaining minority set of cases, the top scoring baseline pose family was correct just 25% of the time. However, the correct choice could be found 50% of the time by looking at all three of the top scoring families. Success rates of 50% approach those observed with cognate docking on difficult benchmarks, but the comparable rates for those studies come from consideration of a *single* top scoring pose instead of the poses from a trio of families.

Inter-target variation

The tremendous variation in docking system performance on target choice has been well documented (e.g. [12, 22]).

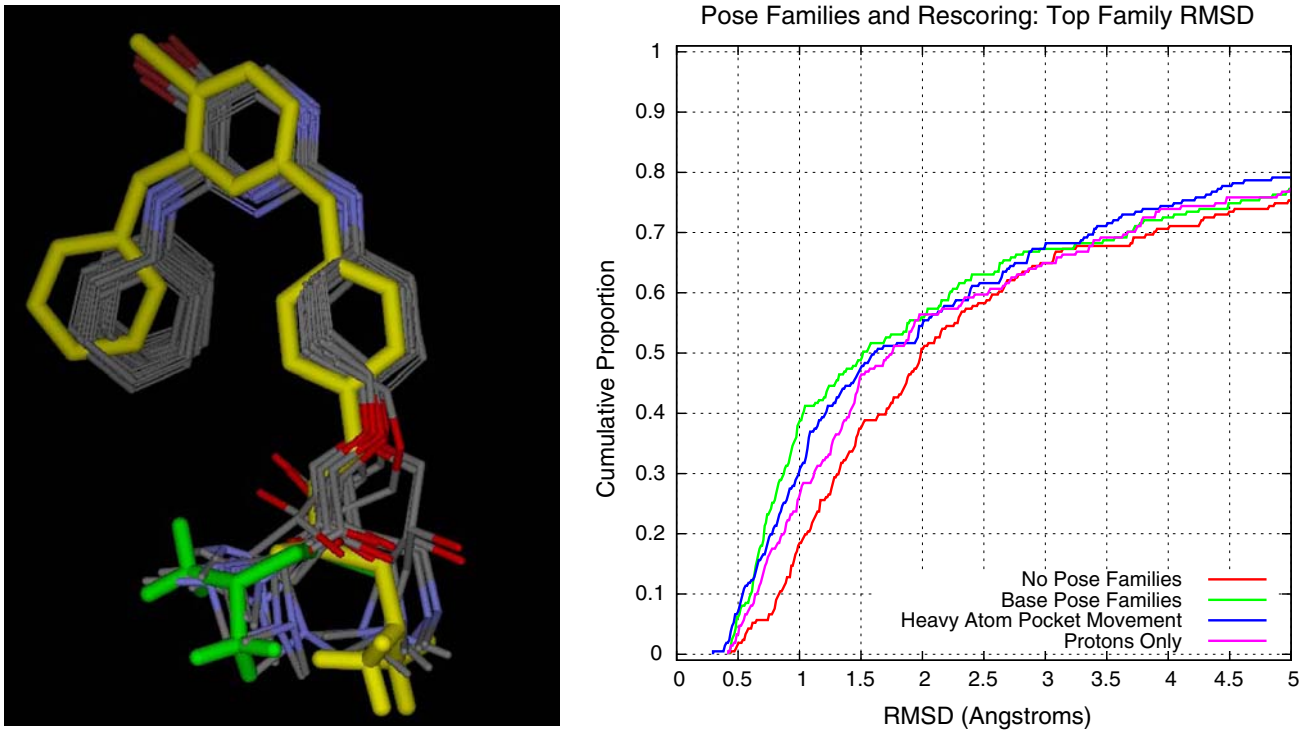

Fig. 9 At *left* is a depiction of the top scoring pose family for a ligand of CDK2. The portion of the ligand that is deep within the pocket (at top) exhibits relatively little variation, but the portion that extends toward solvent exhibits a variety of reasonable orientations. The crystallographic determination yielded two alternative conformations (shown in *yellow* and *green*), which are spanned by the pose family. At *right* is shown the comparison of using purely the top scoring pose of a ligand (*red line*) compared with using the top pose families from either the initial docking (*green line*), the result of post-docking pocket optimization with all protein atoms (*blue line*), or post-docking pocket optimization with protons only (*purple line*). The use of pose families makes only a nominal improvement at the 2 Å level, but the physical depiction of pose variation is likely to be useful, as in this case where an accurate depiction of mobility is made

Table 1 shows the performance of the multi-structure pocket-adaptation protocol for Surflex-Dock over the set of eight targets. Average success rates for multi-structure docking with no rescoring or pose family computation ranged from 40% for thrombin to 79% for estrogen receptor, with the mean being 61%. Weaker performance for thrombin was primarily due to extreme ligand flexibility in many cases, along with the previously mentioned issues of P1 pocket element variability and the presence of ligands that require metal chelation not present in the protein structures used for docking. CDK2 represents a genuinely difficult case, since the protein motions captured with the five protein conformational snapshots clearly do not encompass finer motions that are important (Fig. 6).

Rescoring with protein pocket adaptation had large effects on individual target performance, but due to small numbers of ligands per target, these were not statistically significant. Interestingly, the largest difference in performance between the two rescoring approaches were between the proteins representing the two poles of relative flexibility, with full pocket adaptation performance better on CDK2 and proton-only adaptation performing better on PDE4b/5a. The aggregate mean performance differences were not significant. However, consideration of *two* pose families (either the original top family and the top family from full pocket adaptation or the former plus that from proton-only pocket adaptation) yielded highly significant performance improvements over performance without any pocket adaptation. In terms of the practical impact on modeling, a requirement to employ judgment given two or three solution sets (and only in the cases where they disagree) does not seem overly burdensome. Note that consideration of the two most probable pose families from the baseline docking (*without* pocket adaptation) yielded performance levels that were not statistically significantly different than those shown in Table 1 for two pose families obtained using pocket adaptation (Orig + Heavy and Orig + Protons). However, pocket adaptation allows the computation of pose family agreement (discussed above) since the protocols employ scoring variations. Also, pocket-adaptation can yield significant changes to protein-ligand interactions and pose family composition (Figs. 8, 9).

Figure 11 shows an example from PDE4b, where the top scoring pose family from proton-only adaptation was correct. In this case, the uncertainty in the placement of the

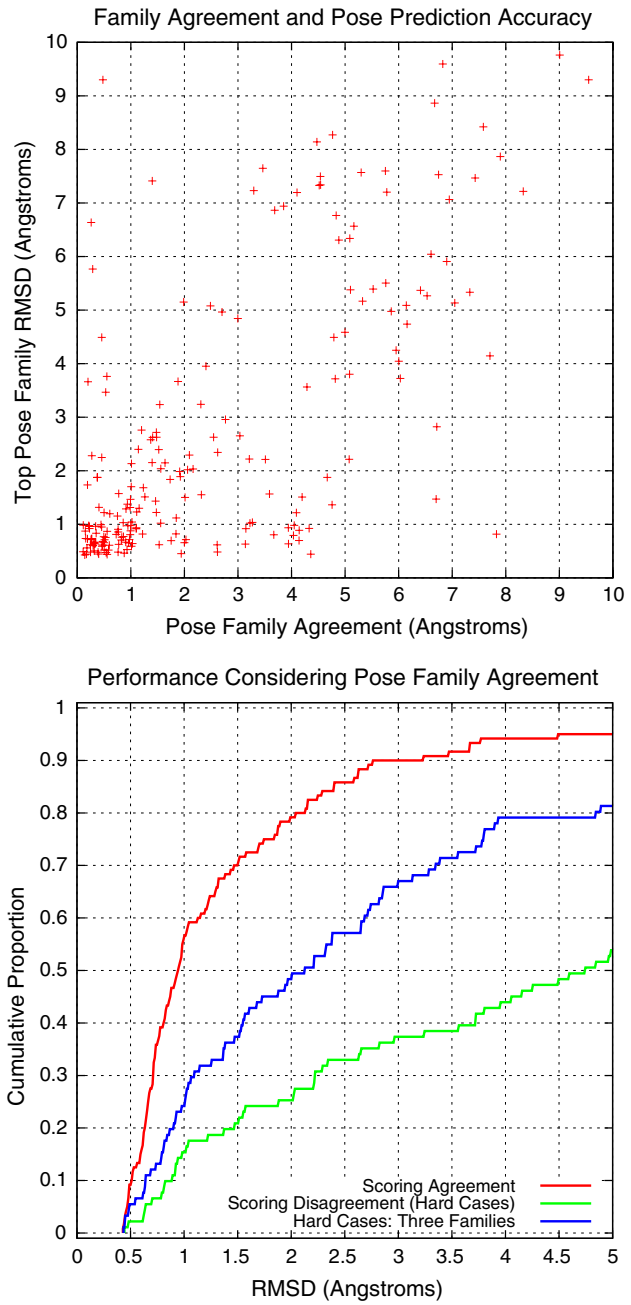

Fig. 10 The *top plot* shows the relationship between pose family agreement (see text) and the accuracy of the top scoring pose family from the non-rescored docking run. There is a very strong relationship (Kendall's Tau 0.45, $p \ll 0.01$ by permutation). The *bottom plot* shows the cumulative histogram of predicted single pose family accuracy for the cases in which pose family agreement is high (*red*) or low (*green*). In the high agreement cases (120/211 or 57%), the expectation is that 80% of the time, the top pose family contains the correct docking. Conversely, for the cases of disagreement (the remaining 43%), the success rate is closer to 25%. However, if we consider the top pose family for each of the three scoring methods (*blue*), our success rate doubles, to 50%. These differences are highly statistically significant (Fisher's exact test on the difference of proportions of success/failure at 2.0 Å rmsd). Note that the high-agreement cases involve ligands that do not differ in flexibility than the low-agreement cases (6.2 vs. 7.4)

chlorophenyl seems warranted in light of the partial density from the crystal structure in that part of the ligand. Fairly subtle protons movements (highlighted in the Figure) were important in proper recognition of the correct pose. In this case, the ligand to be docked shares some commonality with the known cognate ligand structures (Fig. 3), but a number of reasonable "flips" are easily confusable, since the core heterocycle is functionalized differently, both in position and content, compared with the nearest known analog. For PDE4b, the proton-only approach appears more reliable, probably due to the a priori fact of relative protein rigidity. The combined force-field within Surflex-Dock in the pocket adaptation protocol is not reliable in this case when moving heavy atoms, adding more noise than signal to the scores.

In the case of estrogen receptor, all three methods worked quite well, with a high level of agreement and with a combined performance of 95% correct pose prediction. Figure 12 shows a typical example for this target, where an antagonist (from 1UOM, shown also in Fig. 3) was the subject of docking. This ligand represents the type of synthetic variation one would encounter in lead optimization exercises, where the antagonist "arm" is among the structures known, but the core structure that binds in the agonist pocket is quite different from the known ligand structures. The all-atom pocket adaptation approach is robust enough to "rescue" the correct pose of the antagonist when bound to an *agonist*-form of the receptor. However, as can be seen in Fig. 12 (right panel), the pocket adaptation, while making room for the ligand, does not even come close to adapting the pocket to the form seen when binding antagonists. The approach taken here will be most successful in cases where the large protein motions are well-represented among a small set of experimentally determined structures. The only ligand that represented a failure was that from 1ZKY, which binds the agonist binding site but has a complex bicyclic structure. In that case, the top pose family from the protons-only rescoring was still within 3.0 Å rmsd, which was the closest solution among all of the dockings returned.

Figure 13 shows the docking of the ligand from 1FPC into the five alternate thrombin structures (see Fig. 3 for 2D structures). The original docking contained the correct solution, but it was ranked a full 2.0 units of pK_d lower than the incorrect solution shown in the left panel. Rescoring using either pocket adaptation method yielded the correct family as top-ranking (middle panel). While the movement of TRP-86 is helpful to accommodate the larger substituent (compared with the cognate ligands), it is likely that inclusion of non-covalent ligand self-interaction, which is part of the pocket-adaptation rescoring procedure, is beneficial for the entire class of thrombin inhibitors with this typical three-part construction.

Table 1 Performance of Surflex-Dock on a target-specific basis

Target	*N* test ligands	Mean rot bonds	Success rates (proportion ≤ 2.0 Å rmsd)						
			Top pose	Original top pose family	Pocket optimization		Two pose families		Three
					Heavy atom	Protons	Orig+heavy	Orig+protons	All 3
ESR1	19	5.9	79	79	74	95	84	95	95
PDE4b/5a	12	5.7	75	75	25	67	83	92	92
MMP8/13	11	10.5	64	82	73	82	82	82	82
MAPK14	20	5.2	60	65	70	60	70	65	70
CDK2	79	5.5	48	53	56	46	65	59	67
F2 (thrombin)	70	8.4	40	46	50	51	60	60	69
Mean	35.2	6.9	60.9	66.6	57.9	66.7	74.0	75.5	79.0

Results are shown using either the single top pose returned from a multi-protein-conformation docking (including use of fragment-based hints), using the top pose family under different rescoring protocols, or using multiple pose families. The differences in success rates among the single pose and single pose family protocols are not statistically significant, either in terms of average success rates or in terms of proportion of success overall. However, use of two pose families or three yields a highly significant improvement compared with using a single pose or single pose family (Fisher's exact test of the difference of proportions of success/failure)

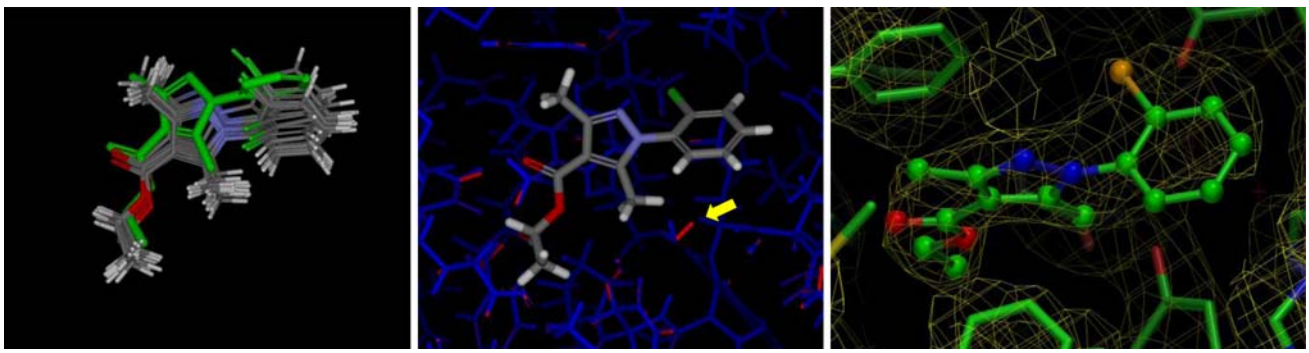

Fig. 11 Cross docking of the ligand from 1Y2H into PDE4b. At *left*, the top pose family from the proton-based pocket refinement (probability 1.00), is shown along with the crystallographic pose (*green*). There is a good deal of uncertainty in the placement of the chlorophenyl, which has an impact on the position of the remainder of the ligand. The *center panel* shows the original protein conformation (*blue*) and the modified one (*red*) that leads to the most dominant ligand pose from the pose family. Reorientation of a hydroxyl proton (TYR233, indicated by an *arrow* in the *middle panel*, at *bottom* on *right*) is critical to allow room for the ligand, and minor movement of a donor proton on GLN443 is also important in yielding correct recognition. The ligand extends well beyond the density in the area of the chlorophenyl (*right*), which suggests that alternative orientations are reasonable to propose

Relationship to other approaches

The work reported here represents a contribution to real-world docking primarily in four ways. First, the approach is computationally tractable, with typical per-ligand computation times of about 30 min. With multi-CPU clusters being common, ligand sets under consideration in lead-optimization exercises can be thoroughly studied with these methods. Second, the benchmark used here contains a small number of pharmaceutically relevant targets, represented each with a small number of conformational snapshots, but the testing was done with a large number of ligands of highly variable structure in many cases. Further, the benchmark itself was not constructed by a methods developer to demonstrate performance of a particular method; rather it was constructed by an independent active modeler in order to measure real-world behavior. Third, the approach offers a way to systematically make use of modeling knowledge in the form of ligand fragments and their key interactions, but to do so in a way that does not lead to undue bias in constraining the prediction space. Fourth, the workflow yields physically intuitive results: related pose families under a small number of scoring conditions that allow for significant protein flexibility including both sidechain and backbone movements.

These results represent very significant practical improvements over single-structure non-cognate docking. Single top-scoring pose family predicted performance averaged 64% (baseline multi-structure docking, heavy-atom pocket optimization, and proton-only pocket optimization). When top pose families agreed, 80% correct prediction was observed. Overall, consideration of the best

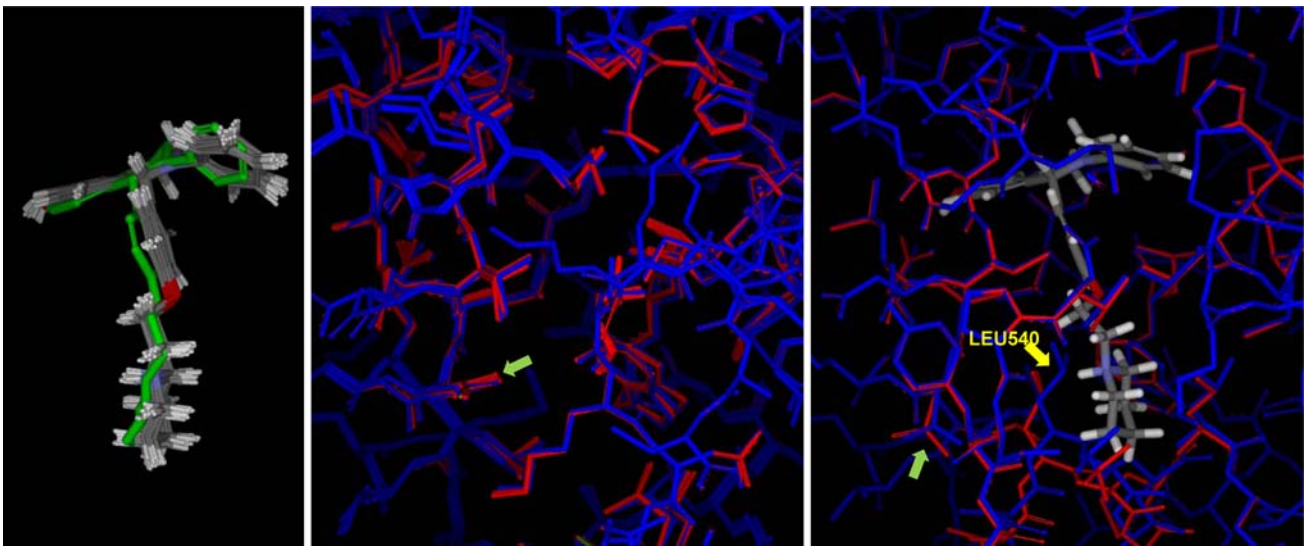

Fig. 12 Cross docking of the ligand from 1UOM into ESR1. In this case, all three scoring methods agreed on the top scoring pose family. At *left*, the crystallographic pose is shown with the pose family from heavy-atom pocket optimization. Only the antagonist structures (1YIM, 1SJ0, and 2ERT) contributed significantly to the pose family shown. In the *middle*, the protein atom movement is shown (*red*), which is minimal in this case. The ligand is relatively similar in structure and binding preference to the three cognate antagonists among the five structures used. At *right*, a pose resulting from docking to an *agonist* structure (1X7R). This pose is reasonable and close to correct, but the protein conformation resulting from heavy atom optimization cannot replicate the wholesale rearrangement of the protein (ASP351 is marked in both panels with a *green arrow*). LEU540 (labeled in the *right panel*) moves so much in the true antagonist-bound form that it does not appear in the *middle depiction*

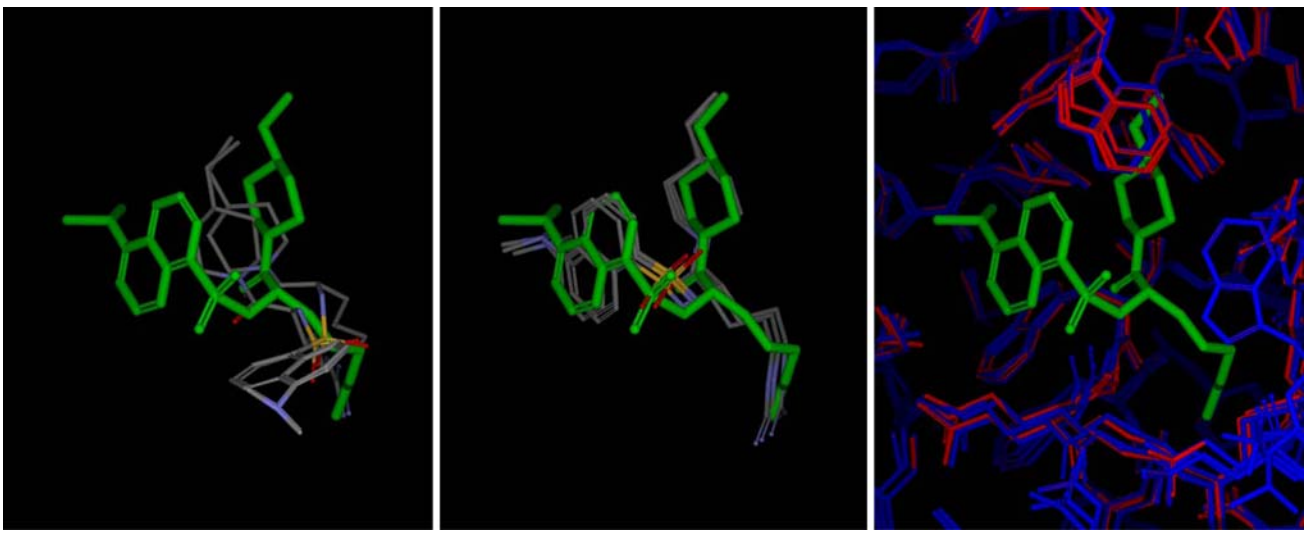

Fig. 13 Cross docking of the ligand from 1FPC into F2. In this case, the original baseline docking yielded an incorrect top pose family, with the guanidinium correctly placed, but with the napthyl substituent significantly misplaced (shown at *left*). However, both methods of rescoring with protein pocket adaptation yielded the correct pose in the top family (*middle panel*). Accommodation of the ethyl-pyridine involves movement of TRP86 when heavy atoms are allowed to move (*right panel*). It is likely that ligand non-covalent self-interaction also contributes toward improved recognition in this and similar cases

prediction from the two top pose families from two scoring methods yielded correct predictions 75% of the time, averaged across the targets tested. *Cognate* docking on the same set yielded 65% success, so the results for cross-docking with this multi-pronged approach are competitive.

This work was positively influenced by much of the work that has been previously reported addressing protein flexibility, particularly including that from the groups of Abagyan, Gilson, Friesner, Goodsell, McCammon, Moitessier, and Shoichet [30–36, 40]. The foregoing work has generally focused on elegant studies of single targets or all-

by-all cross dockings with small total numbers of ligands (generally 25 or less). The present work has made use of a very large testing set of realistic construction (211 test ligands for eight total targets, with five starting protein conformations per target). It is difficult to make sensible comparisons in terms of performance levels since the studies are so different, but the results shown here are transparently relatable to real-world use scenarios, and performance levels approach those seen in multiple studies on "hard" benchmarks for *cognate* docking. Among the prior reported methods in which true protein flexibility has been explored, processing times spanned multiple hours for single ligands, compared with the 30-min timings typical in this study (for an initial multi-structure docking, rescoring with all-atom protein pocket adaptation, and pose family generation for the baseline and rescored poses).

Conclusions

In recently published work that laid the computational foundation for the work presented here, the use of protein coordinate optimization in the presence of cognate ligands was shown to yield significant bias effects in nominal performance for pose prediction in cognate docking [10]. In that paper, the following hope was expressed:

> … that significant improvements, particularly in docking accuracy, should be possible and should not necessarily require combinatorial exploration of protein configurational space simultaneously with ligand configurational space. It may be possible to employ local optimization of protein active site atoms, after docking, to obtain these benefits without incurring a burdensome computational cost.

The work reported here demonstrates a step along the path, with clear improvements in docking accuracy as a result of considering such pocket adaptation. There remains much to be done, however. While the computational cost is not overwhelming, a goal of closer to 5–10 min per ligand seems attainable at the performance levels observed here. More importantly, as we have seen in work on scoring function tuning [11], there is an opportunity to improve the overall scoring regime, possibly on a target-specific basis. Tuning of the non-covalent Surflex-Dock scoring function *with* protein movement is expected to yield stiffer clashing penalties along with sharpened terms for both hydrophobic and polar interactions. Since the number of parameters in the covalent protein force-field is relatively small, those parameters should also be amenable to tuning within the multiple-instance paradigm used previously. There is no reason to believe that the *particular* parameters chosen by Mayo et al. [41] for the DREIDING force-field to optimally predict small molecule geometries should be particularly well-suited to scoring the blended interactions within protein/ligand complexes, as has been done here. This represents a significant opportunity, but also a challenge, since parameter optimization must account for the changes to optimal configurations, and configurational optimization that includes the protein pockets is computationally somewhat costly.

Nonetheless, the methods leading to the pose-prediction performance reported here on a large, realistic, and pharmaceutically relevant cross-docking benchmark should be of use to real-world modelers.

Acknowledgments The author gratefully acknowledges NIH for partial funding of the work (grant GM070481). He is especially grateful to Dr. Jeffrey Sutherland for sharing his extensively curated cross-docking data set and to the authors of Ref-25 for making the Astex85 Set available. He is also grateful for comments on the manuscript and discussion from Dr. Ann Cleves. Dr. Jain has a financial interest in BioPharmics LLC, a biotechnology company whose main focus is in the development of methods for computational modeling in drug discovery. Tripos Inc., has exclusive commercial distribution rights for Surflex-Dock, licensed from BioPharmics LLC.

References

1. Kuntz ID, Blaney JM, Oatley SJ, Langridge R, Ferrin TE (1982) A geometric approach to macromolecule-ligand interactions. J Mol Biol 161(2):269–288
2. Welch W, Ruppert J, Jain AN (1996) Hammerhead: fast, fully automated docking of flexible ligands to protein binding sites. Chem Biol 3(6):449–462
3. Rarey M, Kramer B, Lengauer T, Klebe G (1996) A fast flexible docking method using an incremental construction algorithm. J Mol Biol 261(3):470–489
4. Jones G, Willett P, Glen RC, Leach AR, Taylor R (1997) Development and validation of a genetic algorithm for flexible docking. J Mol Biol 267(3):727–748
5. Goodsell DS, Morris GM, Olson AJ (1996) Automated docking of flexible ligands: applications of AutoDock. J Mol Recognit 9(1):1–5
6. Bohm HJ (1994) The development of a simple empirical scoring function to estimate the binding constant for a protein-ligand complex of known three-dimensional structure. J Comput Aided Mol Des 8(3):243–256
7. Jain AN (1996) Scoring noncovalent protein-ligand interactions: a continuous differentiable function tuned to compute binding affinities. J Comput Aided Mol Des 10(5):427–440
8. Eldridge MD, Murray CW, Auton TR, Paolini GV, Mee RP (1997) Empirical scoring functions: I. The development of a fast empirical scoring function to estimate the binding affinity of ligands in receptor complexes. J Comput Aided Mol Des 11(5):425–445
9. Jain AN (2003) Surflex: fully automatic flexible molecular docking using a molecular similarity-based search engine. J Med Chem 46(4):499–511

10. Jain AN (2007) Surflex-Dock 2.1: robust performance from ligand energetic modeling, ring flexibility, and knowledge-based search. J Comput Aided Mol Des 21(5):281–306
11. Pham TA, Jain AN (2008) Customizing scoring functions for docking. J Comput Aided Mol Des 22(5):269–286
12. Warren GL, Andrews CW, Capelli AM, Clarke B, LaLonde J, Lambert MH, Lindvall M, Nevins N, Semus SF, Senger S, Tedesco G, Wall ID, Woolven JM, Peishoff CE, Head MS (2006) A critical assessment of docking programs and scoring functions. J Med Chem 49(20):5912–5931
13. Jain AN (2004) Ligand-based structural hypotheses for virtual screening. J Med Chem 47(4):947–961
14. Gilson MK, Given JA, Head MS (1997) A new class of models for computing receptor-ligand binding affinities. Chem Biol 4(2):87–92
15. Bissantz C, Folkers G, Rognan D (2000) Protein-based virtual screening of chemical databases. 1. Evaluation of different docking/scoring combinations. J Med Chem 43(25):4759–4767
16. Huang N, Shoichet BK, Irwin JJ (2006) Benchmarking sets for molecular docking. J Med Chem 49(23):6789–6801
17. Miteva MA, Lee WH, Montes MO, Villoutreix BO (2005) Fast structure-based virtual ligand screening combining FRED, DOCK, and Surflex. J Med Chem 48(19):6012–6022
18. Perola E, Walters WP, Charifson PS (2004) A detailed comparison of current docking and scoring methods on systems of pharmaceutical relevance. Proteins 56(2):235–249
19. Kellenberger E, Rodrigo J, Muller P, Rognan D (2004) Comparative evaluation of eight docking tools for docking and virtual screening accuracy. Proteins 57(2):225–242
20. Irwin JJ (2008) Community benchmarks for virtual screening. J Comput Aided Mol Des 22(3–4):193–199
21. Liebeschuetz JW (2008) Evaluating docking programs: keeping the playing field level. J Comput Aided Mol Des 22(3–4):229–238
22. Nicholls A (2008) What do we know and when do we know it? J Comput Aided Mol Des 22(3–4):239–255
23. Jain AN (2008) Bias, reporting, and sharing: computational evaluations of docking methods. J Comput Aided Mol Des 22(3–4):201–212
24. Jain AN, Nicholls A (2008) Recommendations for evaluation of computational methods. J Comput Aided Mol Des 22(3–4):133–139
25. Hartshorn MJ, Verdonk ML, Chessari G, Brewerton SC, Mooij WT, Mortenson PN, Murray CW (2007) Diverse, high-quality test set for the validation of protein-ligand docking performance. J Med Chem 50(4):726–741
26. Pham TA, Jain AN (2006) Parameter estimation for scoring protein-ligand interactions using negative training data. J Med Chem 49(20):5856–5868
27. Jain AN (2006) Scoring functions for protein-ligand docking. Curr Protein Pept Sci 7(5):407–420
28. Dietterich TG, Lathrop RH, Lozano-Perez T (1997) Solving the multiple instance problem with axis-parallel rectangles. Artif Intell 89(1–2):31–71
29. Jain AN, Dietterich TG, Lathrop RH, Chapman D, Critchlow RE, Bauer BE, Webster TA, Lozano-Perez T (1994) A shape-based machine learning tool for drug design. J Comput Aided Mol Des 8(6):635–652
30. Lin J-H, Perryman AL, Schames JR, McCammon JA (2002) Computational drug design accommodating receptor flexibility: the relaxed complex scheme. J Am Chem Soc 124(20):5632–5633
31. Amaro RE, Baron R, McCammon JA (2008) An improved relaxed complex scheme for receptor flexibility in computer-aided drug design. J Comput Aided Mol Des 22(9):693–705
32. Osterberg F, Morris GM, Sanner MF, Olson AJ, Goodsell DS (2002) Automated docking to multiple target structures: incorporation of protein mobility and structural water heterogeneity in AutoDock. Proteins 46(1):34–40
33. Kairys V, Gilson MK (2002) Enhanced docking with the mining minima optimizer: acceleration and side-chain flexibility. J Comput Chem 23(16):1656–1670
34. Cavasotto CN, Abagyan RA (2004) Protein flexibility in ligand docking and virtual screening to protein kinases. J Mol Biol 337(1):209–225
35. Wei BQ, Weaver LH, Ferrari AM, Matthews BW, Shoichet BK (2004) Testing a flexible-receptor docking algorithm in a model binding site. J Mol Biol 337(5):1161–1182
36. Sherman W, Day T, Jacobson MP, Friesner RA, Farid R (2006) Novel procedure for modeling ligand/receptor induced fit effects. J Med Chem 49(2):534–553
37. Metwally E, Shepphird JK (2007) Surflex-Dock: effects of protomol generation and fragment matching on docking results. ACS Fall 2007 symposium (p. (oral presentation)). American Chemical Society, Boston
38. Corbeil CR, Englebienne P, Moitessier N (2007) Docking ligands into flexible and solvated macromolecules. 1. Development and validation of FITTED 1.0. J Chem Inf Model 47(2):435–449
39. Verdonk ML, Mortenson PN, Hall RJ, Hartshorn MJ, Murray CW (2008) Protein-ligand docking against non-native protein conformers. J Chem Inf Model 48(11):2214–2225
40. Sutherland JJ, Nandigam RK, Erickson JA, Vieth M (2007) Lessons in molecular recognition. 2. Assessing and improving cross-docking accuracy. J Chem Inf Model 47(6):2293–2302
41. Mayo SL (1990) DREIDING: a generic force field for molecular simulations. J Phys Chem 94(26):8897–8909